DOCUMENT RESUME

ED 074 081 TM 002 443

AUTHOR Howell, John F.; Games, Paul A.
TITLE The Effects of Variance Heterogeneity on Simultaneous Multiple Comparison Procedures with Equal Sample Size.
PUB DATE Feb 73
NOTE 15p.; Paper presented at the American Educational Research Association Convention, February 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Analysis of Variance; *Comparative Statistics; *Computer Programs; *Statistical Analysis; Technical Reports; *Test Reviews

ABSTRACT

The two purposes of this investigation were to study the effects of variance heterogeneity on three selected multiple comparison procedures and to determine if either of two nonstandard methods would be superior to the conventional methods based on mean square within. The three procedures studied were the Wholly Significant Difference Test (WSD), The "S" test, and a simple multiple "t" test (MTT) procedure. The investigation was a computer simulation consisting of 1000 experiments with four independent samples of five data points. Six pairwise contrasts were considered. The four variance conditions (VC) constituted one factor of the design. Each of the six contrasts were tested using three methods. The three methods constituted a second factor in the two-factor design with VC crossed with method. Results are tabulated and discussed. (DB)

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 074081

TM 002 443

THE EFFECTS OF VARIANCE HETEROGENEITY ON SIMULTANEOUS MULTIPLE COMPARISON PROCEDURES WITH EQUAL SAMPLE SIZE

John F. Howell and Paul A. Games

The Pennsylvania State University

Paper presented at the American Educational Research Association Convention, February, 1973.

THE EFFECTS OF VARIANCE HETEROGENEITY ON SIMULTANEOUS MULTIPLE COMPARISON PROCEDURES WITH EQUAL SAMPLE SIZE

John F. Howell[1] and Paul A. Games

The Pennsylvania State University

Educational and psychological researchers often deal with groups that tend to be heterogeneous in variability. The purpose of this investigation was to study the effects of variance heterogeneity on three selected multiple comparison procedures, and to determine if either of two non-standard methods would be superior to the conventional methods based on mean square within.

The three procedures studied were the Wholly Significant Difference Test (WSD) developed by Tukey (1953), the S test presented by Scheffé (1953), and a simple multiple t test (MTT) procedure. These three were selected because each controls a distinctly different Type I error rate. The WSD controls the familywise Type I error rate (FWI), the MTT controls the contrast significance rate, and the S test controls the risk of finding at least one significant contrast in a set of all possible contrasts. Games (1971a, 1971b) showed that all three procedures use the same statistic and hence any differences between the three procedures are due to the use of different critical values (CV).

---

[1]Now with the Springfield Public Schools, Springfield, Massachusetts.

Computer time that made this study possible was supplied by the Pennsylvania State University Computation Center.

Method

This investigation was a computer simulation consisting of 1000 experiments with four independent samples ($k = 4$) of five data points ($n = 5$). Six pairwise contrasts were considered. Each data point, selected at random from a population of 10,000 normal deviates, was scaled by multiplying it by a specifically chosen constant to create variance ratios of 4:4:4:4 (VC-1), 1:3:5:7 (VC-2), 1:1:7:7 (VC-3), and 1:1:1:13 (VC-4). The four variance conditions (VC) constituted one factor of the design.

Each of the six contrasts were tested using three methods. Method MSW consisted of the conventional test using the square root of two Mean Square Within divided by n as the denominator of a t with df =16. Method t used the standard error of the common t test with df =8. Method BF used the Behrens-Fisher t' statistic defined as

$$(\bar{X}_i - \bar{X}_j) \Big/ \sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}$$ with df given by the Welch solution,

(Winer, 1962, p.37) recommended by Scheffé (1970) and Wang (1971). The three methods constitute a second factor in the two-factor design with VC crossed with method.

Conditions when the null hypothesis is false were created by adding a constant to each data point. A uniform distribution was used spreading the population means equally apart. The distance between the means was calculated by a formula modified from Games and Lucas (1966, p.317).

Five dependent measures were recorded simultaneously: the average contrast Type I rate ($\bar{P}$), the familywise Type I rate (FWI), specific contrasts $\bar{X}_1 - \bar{X}_4$, $\bar{X}_1 - \bar{X}_2$, and $\bar{X}_3 - \bar{X}_4$.

A two factor within-replications analysis of variance (AOV) was performed for each of the five dependent variables, for the null hypothesis condition, both VC and SE being repeated measure factors. The analysis was performed on the IBM 360/67 using the library routine ANOVR (Pitz, 1968).

Results

Per comparison rates are presented in Table I, one for each method. With the exception of the .084 value for the MSW under VC-4, all of the deviations of $\bar{P}$ from .05 are relatively minor. Furthermore, as expected, the use of MSW produces the greatest power under the homogeneous condition (VC-1). Only under the most extreme variance condition (VC-4) did the use of MSW produce an inflated Type I error rate (and a lower power curve). Therefore, using $\bar{P}$ as a dependent measure suggests the universal use of MSW in all but extremely heterogeneous conditions. The robustness of the common multiple comparison solution can be defended on the basis of $\bar{P}$ data, at least for the equal n case.

Table I

Obtained Proportions of Rejections for the Multiple t Test Using $\alpha = .05$, $n = 5$

		$\phi=0$			$\phi=1.0$			$\phi=1.5$			$\phi=2.0$		
	VC	MSW	t	BF	MSW	t	BF	MSW	t	BF	MSW	t	BF
FWI	1	.185	.214	.176	.571	.546	.507	.842	.809	.784	.971	.942	.923
	2	.195	.226	.181	.576	.608	.529	.839	.870	.799	.968	.964	.935
	3	.219	.232	.166	.593	.609	.528	.833	.869	.801	.966	.966	.946
	4	.233	.224	.164	.550	.713	.651	.791	.955	.932	.940	.997	.996
Theoretical		.189	.175										
$\bar{P}$	1	.048	.050	.040	.199	.178	.158	.345	.311	.283	.491	.459	.419
	2	.054	.058	.045	.204	.192	.158	.346	.360	.310	.489	.502	.451
	3	.064	.064	.045	.213	.210	.168	.348	.357	.302	.487	.507	.448
	4	.084	.064	.045	.207	.257	.214	.341	.417	.361	.487	.567	.508
Theoretical		.050	.050										
$\bar{X}_1-\bar{X}_4$	1	.055	.049	.042	.461	.397	.375	.778	.714	.673	.946	.911	.882
	2	.056	.060	.041	.464	.424	.336	.770	.707	.619	.939	.906	.829
	3	.063	.060	.041	.467	.424	.337	.759	.707	.619	.939	.906	.829
	4	.162	.071	.044	.487	.280	.217	.722	.517	.392	.888	.709	.615
$\bar{X}_1-\bar{X}_2$	1	.047	.042	.032	.074	.071	.060	.140	.115	.095	.216	.193	.167
	2	.007	.045	.038	.031	.106	.087	.071	.231	.188	.160	.363	.321
	3	.000	.042	.032	.012	.193	.167	.044	.396	.361	.120	.620	.574
	4	.004	.042	.032	.022	.193	.167	.067	.396	.361	.163	.620	.574
$\bar{X}_3-\bar{X}_4$	1	.052	.054	.044	.092	.092	.077	.162	.142	.115	.244	.227	.194
	2	.105	.054	.046	.165	.077	.068	.215	.109	.091	.303	.166	.132
	3	.140	.054	.044	.197	.076	.058	.255	.103	.089	.323	.144	.155
	4	.161	.078	.050	.211	.102	.067	.258	.136	.091	.340	.180	.133

However, the above analysis is superficial since control of $\bar{P}$ does not imply adequate control over the individual rates from which the average was obtained. Three specific contrasts were isolated ($\bar{X}_1 - \bar{X}_4$, $\bar{X}_1 - \bar{X}_2$, and $\bar{X}_3 - \bar{X}_4$) as the most interesting in terms of the manner in which the heterogeneous conditions were established.

As an example, assume a researcher tested the difference between the means of the third and fourth groups ($\bar{X}_3 - \bar{X}_4$) and that between the first and second groups ($\bar{X}_1 - \bar{X}_2$) on a priori experimental grounds, using the pooled variance estimate MSW. The values for the two contrasts that he selected (from Table I) are presented below. In every heterogeneous variance condition, $\bar{P}$ underestimates the rate for the $\bar{X}_3 - \bar{X}_4$ contrast and overestimates it for the $\bar{X}_1 - \bar{X}_2$ contrast.

Variances	$\bar{P}$	$\bar{X}_3 - \bar{X}_4$	$\bar{X}_1 - \bar{X}_2$
4:4:4:4	.048	.052	.047
1:3:5:7	.054	.105	.007
1:1:7:7	.064	.140	.000
1:1:1:13	.084	.161	.004

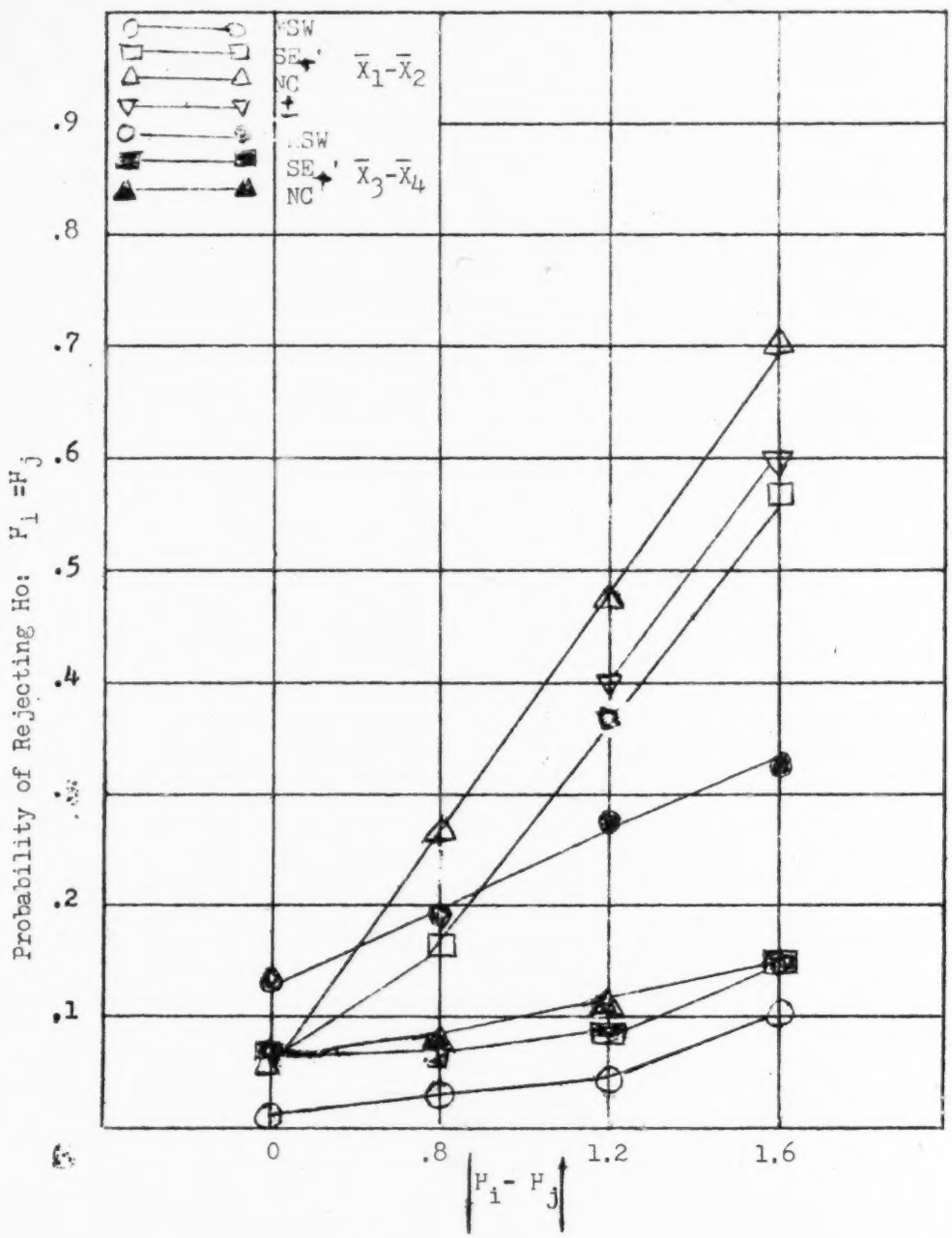
The $\bar{P}$ results hide the great inaccuracies on individual contrasts. $\bar{P}$ =.054 may be obtained as the average of two contrasts with P(EI) =.14, two with P(EI) =.000, and two with P(EI) =.063 (as suggested by the results on VC-3). However, if the researcher above tested his means using the t test, he would achieve more uniform control. Using the Behrens-Fisher method, results in P(EI)≈.05 regardless of the individual contrast selected as shown below.

Variances	$\bar{P}$	$\bar{X}_3 - \bar{X}_4$	$\bar{X}_1 - \bar{X}_2$
4:4:4:4	.040	.044	.032
1:3:5:7	.045	.046	.038
1:1:7:7	.045	.044	.032
1:1:1:13	.045	.050	.032

An AOV was performed using the three individual contrasts as a third within-replications factor (CON) with VC and method. As expected, the analysis resulted in a significant CON x VC x method interaction ($F = 72.9$, df $=12,38$; $p < .001$). The CON x VC interaction was different for each method. A two factor repeated measures design with CON and VC the two factors was conducted for each method. When using MSW there was a wide difference in P(EI) in the three contrasts as the variance condition changed. There was a significant CON x VC interaction ($F=148.9$; df$=6,18$; $p < .001$). Using the $\underline{t}$ method produced considerable improvement but still resulted in a significant CON x VC interaction ($F=6.324$; df$=6,18$; $p < .001$). However, for the BF method no significant differences were noted for either factor or the interaction. This result suggests that using the BF method when violation of the homogeneity of variance assumption is suspected will result in the over-all stability of Type I error around α.

Figure 1 illustrates the power curves for two contrasts, $\bar{X}_1-\bar{X}_2$ and $\bar{X}_3-\bar{X}_4$ when the variance ratio 1:1:7:7 (VC-3) was used. The theoretical normal curve power solutions for these two contrasts have also been inserted. The normal curve solutions are available since the population variances are known. These power curves are higher than similar power curves using the proper $\underline{t}$ distribution. Some points for the $\underline{t}$ solutions are not available due to limitations in available tables. The power curves for the contrasts using MSW greatly diverge from the theoretical

Figure 1. Power curves for $\bar{X}_1-\bar{X}_2$, $\bar{X}_3-\bar{X}_4$, theoretical normal curves (NC) and theoretical t curves, ($\pm$) under variance condition 1:1:7:7.



curves using the known population variances. The obtained curve on $\bar{X}_1 - \bar{X}_2$ remained conservative throughout when MSW was used. When the BF was used the power rose to .57 and roughly paralleled the theoretical curve. Under these conditions it is evident that the power curves using MSW are <u>not</u> acceptable. The fact that the two values of P(EI) average to .07 is not evidence that the MIT or any other technique based on MSW is robust to heterogeneous variances.

The Familywise Rate

The familywise Type I significance rate (FWI) is the proportion of sets of contrasts that contain at least one significant result. While $\bar{P}$ makes little sense when the individual rates differ, FWI is meaningful regardless of the equality of individual rates. The maximum value of any individual rate establishes the minimum value of the FWI. The maximum value (.162), found under the individual rate $\bar{X}_1 - \bar{X}_4$ for MIT using MSW, is the lowest possible value for FWI under this condition. In general, the more contrasts there are in a set the greater FWI becomes. The Tukey WSD was designed to control the FWI for such a set, i.e., the selection of α for the WSD specifies the theoretical probability of at least one Type I error for the set of contrasts when all of the assumptions are met.

Table 2

Obtained Proportion of Rejections for the Tukey WSD Test using $\alpha = .05$, $n = 5$

		$\phi = 0$			$\phi = 1.0$			$\phi = 1.5$			$\phi = 2.0$		
	VC	MSW	t	BF	MSW	t	BF	MSW	t	BF	MSW	t	BF
FWI	1	.044	.054	.029	.303	.299	.242	.594	.523	.469	.851	.784	.724
	2	.066	.062	.048	.299	.345	.270	.599	.609	.487	.846	.841	.742
	3	.077	.067	.047	.310	.341	.268	.609	.601	.476	.838	.847	.739
	4	.108	.066	.038	.327	.452	.373	.593	.764	.687	.786	.961	.927
Theoretical		.050	.050										
$\overline{P}$	1	.009	.010	.005	.082	.074	.058	.186	.152	.128	.310	.273	.231
	2	.015	.013	.010	.091	.087	.066	.199	.193	.150	.317	.316	.251
	3	.019	.016	.011	.096	.095	.069	.202	.186	.135	.322	.309	.240
	4	.036	.018	.010	.108	.123	.093	.205	.234	.198	.316	.376	.305
Theoretical		.011	.013										

When MSW was used, as in the conventional form of the WSD test (Miller, 1966, Winer, 1962), heterogeneous variances increased the FWI above the .05 level. When the $\underline{t}$ method was used there was only a slight increase in FWI over the four variance conditions. When the BF was used the FWI values were conservative with one value falling significantly below the theoretical .05 level.

Taking the one minus the probability of rejecting at least one contrast when the null hypothesis is false as Familywise Power, the MSW method had higher power than the other two methods in the homogeneous condition. As the variance condition became more heterogeneous, MSW lost its superiority in terms of FWI power, but became inferior only in the extreme heterogeneous condition.

The Scheffé $\underline{S}$ test was designed to control FWI on a set of all possible contrasts. Just as $\overline{P}$ is less than α for the WSD, so the FWI is less than α for the $\underline{S}$ test. Otherwise, the results of the FWI analysis for Scheffé were consistent with those of the WSD. The same trends were found but with lower over all proportions of rejection of Ho.

Discussion

The results above indicate that when variance heterogeneity exists, using a pooled error term as in the MSW method is inappropriate. For various individual contrasts, P(EI) will be inflated while for other contrasts P(EI) will approach zero. The use of MSW produces major distortions in many of the individual contrast power curves. These results will not be alleviated by merely increasing the common sample size.

The results also indicate that the use of a non-pooled error term, as in the t or BF methods provides improved control of P(EI) for all contrasts and an acceptable power curve for all contrasts.

The decision to use MSW or the other methods involves risk either way. The use of MSW when inappropriate risks uncontrolled P(EI) and misleading power curves. Not using MSW under homogeneous variance conditions risks only a slight depression in P(EI) and a comparatively slight uniform loss in power. If the universal use of one method is desired, then that method should be the Behrens-Fisher method.

Testing for variance heterogeneity or extensive experience with the variables might provide evidence for deciding on a method. However, caution should be used in testing variance homogeneity since many tests are sensitive to violations of assumptions regarding the form of the population. (Box, 1953; Games, Winkler and Probert, In Press).

The FWI results show that the use of MSW is <u>not</u> as disadvantageous on this overall index as it is with respect to individual comparisons when $n_i = n_j = n$. The WSD using MSW is as robust to heterogeneous variances as is the analysis of variance. FWI will be somewhat inflated with heterogeneous variances and equal n, but the inflation is at least limited to $2\alpha_T$ or $3\alpha_T$ as a maximum. Unfortunately, this is not a great deal of consolation. The overall control of FWI is achieved by using a conservative critical value that substantially reduces power on each contrast. The phenomenon still exists that various individual contrasts are being tested, often with an inappropriate error term, and that the risk of Type II errors will often be close to 1.0 for many substantial contrasts.

Fortunately, the same technique that improves control of P(EI) on each contrast also improves the robustness of FWI. The BF method

applied to each mean difference yields an FWI that is somewhat conservative when the homogeneous variance condition exists (.038) but which never rises above the theoretical value of .05 even if the assumption is not true. This is in contrast to both the MSW and $\underline{t}$ methods. The BF method may be recommended for the control of either P(EI) or FWI, although the researcher will experience some decrease in power when the homogeneous variance condition is true. Historically, the major disadvantage of using either the $\underline{t}$ or BF methods has been the increased amount of calculation necessary. With the now widespread use of computers, either method can be incorporated into general purpose programs.

With equal n's, as used in this study, the computed $\underline{t}$ and BF statistics will be identical and the only difference between the two methods is in the critical values used. The critical value of t_o is from the $\underline{t}$ sampling distribution with df =2n - 2. $[\underline{t}(\alpha/2, 2n - 2)]$. The critical value of t'_o varies from this value (as a lower limit) to an upper limit of $\underline{t}$ $(\alpha/2, n-1)$. The actual df specified by the Welch solution varies from sample to sample. Thus the only distinction between $\underline{t}$ and BF in this study is in the fact that the BF solution may have a larger critical value that overcomes the slight positive bias in the t statistic when the homogeneity of variance assumption is violated (Box, 1953). As the sample size increases, the difference between t $(\alpha/2, 2n - 2)$ and t $(\alpha/2, n - 1)$ decreases and the results for the t and BF methods would be more similar with both P(EI) and FWI approaching their theoretical values. However, in a computer solution it is appropriate to use the best method for any unspecified n. The Behrens-Fisher method with the Welch solution for critical values is best for the small n situation, and hence is recommended.

References

Box, G. E. P. Non-normality and tests of variance. Biometrika, 1953, 40, 318-335.

Dixon, W. J. & Massey, F. J. Introduction to Statistical Analysis. New York: McGraw-Hill, 1957.

Games, P. A., Winkler, H. B. & Probert, D.A. Robust tests for homogeneity of variance. Educational and Psychological Measurement, (in press).

Games, P. A. The inverse relation between the risks of Type I and Type II errors and suggestions for the unequal n case in multiple comparisons. Psychological Bulletin, 1971a, 75, 97-102.

Games, P. Multiple comparisons of means. American Educational Research Journal, 1971b, 8, 531-565.

Games, P. A. & Lucas, P. A. Power of the analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 1966, 6, 311-327.

Harter, H., Clemm, D., & Guthrie, E. The probability integrals of the range and studentized range. Probability integral and percentage points of the studentized range; critical values for Duncan's new multiple range test. Wright Air Development Center Technical Report 58-484, 1959, Vol II, ASTIA Document No. AD231733.

Li, J. Statistical Interence I. Ann Arbor, Michigan; Edwards Brothers, Inc., 1964.

Miller, R. G. Jr. Simultaneous Statistical Inference. New York: McGraw-Hill, 1966.

Pitz, G. F. Computer analysis of variance program ANOVR, Southern Illinois University, 1968.

Scheffé, H.A. A method of judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, H. Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Association, 1970, 65, 1501-1508.

Tukey, J. W. The problem of multiple comparisons. Unpublished manuscript, Princeton University, 1953.

Wang, Y. Y. Probabilities of the Type I errors of the Welch tests for the Behrens-Fisher problem. Journal of the American Statistical Association, 1971, 605-608.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.